Doi: 10.21608/JHPEI.2025.372933.1045

ORIGINAL ARTICLE **Open Access**

Evaluate the impact of the virtual patient simulator on medical students' clinical reasoning skills through a script concordance test.

Marwa Ahmed El Naggar [1,2] , Fatima Altahir [3]

[1] Medical Education Unit, Community and Family Medicine Department, College of Medicine, Jouf University Sakaka, Kingdom of Saudi Arabia.

[2] Medical Education Department, Faculty of Medicine, Suez Canal University, Egypt.

[3] Department of Internal Medicine, Faculty of Medicine and Health Sciences, Omdurman Islamic University, Khartoum, Sudan.

Abstract

Introduction:

Clinical reasoning is a critical skill in medical practice, yet traditional teaching methods often struggle to develop this competency effectively. Virtual patient simulators offer a promising solution by providing realistic, interactive clinical scenarios for students to practice decision-making in a safe environment. This study aimed to measure the effect of a virtual patient simulator (VPS) on medical students' clinical reasoning skills using the Script Concordance Test (SCT).

Methods:

A quasi-experimental, posttest-only control group design was employed. Fourth-year medical students at Jouf University were randomly assigned to either an intervention group (n=46), which used the In Simu Patient simulator, or a control group (n=46), which received traditional teaching methods. Both groups completed an SCT post-intervention, and their scores were compared to those of an expert panel (n=12). The SCT assessed diagnostic, investigative, and treatment-related reasoning across 20 internal medicine cases. Cronbach's Alpha was calculated to evaluate the reliability of the SCT.

Results:

The intervention group demonstrated significantly higher SCT scores than the control group, particularly in diagnostic and treatment-related questions ($p < 0.05$). The SCT showed good reliability (Cronbach's Alpha = 0.85). Post-hoc analysis revealed significant differences between the intervention group and experts in several cases, indicating areas for further improvement. The control group also showed deviations from expert reasoning, highlighting the limitations of traditional teaching methods.

Conclusion:

Virtual patient simulators significantly enhance medical students' clinical reasoning skills, as measured by the SCT. The findings support the integration of VPS into medical curricula to bridge the gap between theoretical knowledge and practical application. Future research should explore long-term outcomes and optimal implementation strategies for simulation-based learning.

Keywords:

Virtual Patient Simulator, Clinical Reasoning Skills, Medical Education, Script Concordance Test, Simulation-Based Learning, Decision-Making in Medicine.

Received: 05-04-2025 Accepted: 18-07-2025

Published Online: July 2025

How to cite this article

El Naggar M.& Altahir F. "Evaluate the impact of the virtual patient simulator on medical students' clinical reasoning skills through a script concordance test." J Health Prof Edu Innov, Vol. 2, No. 2, July 2025, pp 47-63.
Doi: 10.21608/jhpei.2025.372933.1045

Address for Correspondence

Marwa Ahmed El Naggar,
Medical Education Unit, Community and Family Medicine Department, College of Medicine, Jouf University Sakaka, Kingdon of Saudi Arabia.
Medical education department, Faculty of Medicine, Suez Canal University, Egypt.
Email: marwanagar@yahoo.com. managgar@ju.edu.sa,
Mobile: +966 598668032

Introduction:

Medical education aims to equip students with higher-order cognitive skills, including clinical reasoning, decision-making, and critical thinking, and essential for effective patient care. Developing and accessing these skills is a complex challenge, as traditional assessment methods often involve subjectivity, significant resource allocation, and logistical challenges. While valuable, common approaches, such as multiple-choice questions (MCQS), essays, standardised patients, and simulation centres, frequently fall short in reliably evaluating clinical reasoning due to their inherent limitations, including dependency on external factors and the potential for bias. (1, 2)

The Script Concordance Test (SCT) offers a novel solution, grounded in the script theory of medical decision-making proposed by Schmidt et al. This theory posits that expertise in clinical reasoning develops through the progressive organization of knowledge into "scripts," enabling healthcare professionals to navigate complex medical scenarios efficiently. SCT leverages this framework to assess alignment between students' reasoning processes and expert clinicians' knowledge, using structured clinical vignettes with Likert-scale responses to objectively measure reasoning patterns. (3, 4)

Virtual patient (VP) simulators enhance clinical reasoning education by replicating real-world medical scenarios in a risk-free, interactive environment. These computer-based programs allow learners to make diagnostic and therapeutic decisions while receiving immediate feedback. Unlike real patients, VPS facilitate deliberate practice, enabling repeated exposure to diverse cases in a controlled and standardised manner. Applications like In Simu Patient offer an innovative platform for students to hone their diagnostic skills, promoting efficient and cost-effective learning while supporting the preparation for critical assessments such as the USMLE. (5, 6, 7)

Simulation-based education, including VP simulators, aligns with the overarching goals of medical education systems to improve patient safety and minimise medical errors. By enabling independent practice, detailed feedback, and the safe exploration of diagnostic pathways, VPs empower students to develop clinical reasoning and problem-solving competencies effectively. These tools also address the challenges posed by increased student numbers and limited opportunities for hands-on training, ensuring equitable and scalable learning experiences. (8, 9)

Research indicates that integrating VPs into the medical curriculum offers significant potential to enhance learning outcomes. In Saudi Arabia, however, empirical studies exploring VP integration remain scarce. A study at Sulaiman Al Rajhi Colleges highlighted the positive reception of VPs among preclinical and clinical students, demonstrating their feasibility and alignment with global findings on the benefits for fostering clinical reasoning skills. Despite their promise, effective implementation strategies for VP-based learning require further investigation to maximize their educational impact. (10, 11, 12)

This study builds on the theoretical foundation of script theory and the proven efficacy of VPs and SCTs in enhancing clinical reasoning. It seeks to assess the impact of integrating VPs into problem-based learning (PBL) on medical students' reasoning capabilities. It addresses the critical need for innovative and scalable methods to develop these competencies. By doing so, the research contributes to a growing body of evidence supporting VP adoption in medical education and provides insights into its effective implementation within the local context. (13, 14, 15)

Aim:

Measure the effect of virtual patient simulator on medical students’ clinical reasoning skills using the script concordance test.

Research Objective:

Investigating the effect of VPS on the development of clinical reasoning skills among medical students, by comparing the performance of students exposed to VPS (intervention group) with those receiving traditional teaching methods (control group).

Designing and validating an SCT tailored to assess clinical reasoning within the internal medicine curriculum.

Rationale for the Study:

Clinical reasoning is a cornerstone of medical practice, enabling physicians to evaluate complex patient scenarios and make informed decisions. Despite its critical importance, traditional teaching methods in medical education often struggle to adequately develop clinical reasoning skills, which require exposure to realistic, dynamic, and complex clinical situations. Simulation-based learning has emerged as a promising approach to bridge this gap.

Virtual patient simulators (VPS) are innovative and interactive educational tools that immerse students in realistic, problem-based scenarios. Students can enhance their diagnostic reasoning, decision-making, and problem-solving skills by engaging with VPS in a safe, controlled environment. However, there is limited evidence regarding the measurable impact of VPS on clinical reasoning skills, particularly in the context of undergraduate medical education in internal medicine.

The Script Concordance Test (SCT) is a validated assessment tool that evaluates clinical reasoning in uncertain and ambiguous scenarios, reflecting real-world clinical practice. By measuring students' reasoning processes and comparing them to expert opinions, SCT provides a nuanced understanding of how learners navigate complex clinical problems. Despite its effectiveness, SCT has not been widely utilized with VPS interventions in undergraduate medical education.

The findings of this study have the potential to enhance medical education by offering evidence-based insights into the effectiveness of VPS and SCT. They could also pave the way for integrating advanced simulation tools in medical curricula to better prepare students for clinical practice.

Main study hypothesis

Alternative hypothesis: "Introducing (InSimuPatient) TM for the experimental group of fourth year medical students in internal medicine course will enhance their clinical reasoning skills compared with a control group using (SCT), Accept alternative hypothesis if ($p < 0.05$), and consequently reject the null. (The desired level of confidence will be 95%).

Methods:

Study design: The type of study was quasi-experimental. The posttest-only control group design is a basic experimental design in which participants are randomly assigned to either receive an intervention or not, and the outcome of interest is measured only once after the intervention takes place to determine its effect. It was applied in this study (a comparative study) (as illustrated below).

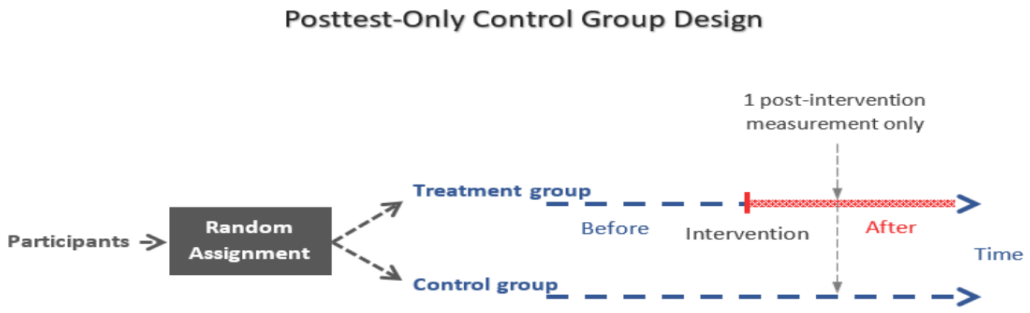

Fig. 1 Posttest-only control group design.

Setting: The College of Medicine, Jouf University, Sakaka. KSA.

Participants: Study population

• Internal medicine course for fourth-year medical students, males and females. Ninety-two students participated in the study in the academic year 2019-2020.

• Internal medicine experts (target population for the validation and development of SCT). Twelve experts participated in the study. The number of experts needed for a Script Concordance Test (SCT) varies depending on the context and the specific objectives of the test. However, research suggests a panel of 10 to 20 experts is ideal to ensure reliable scoring and robust results. A study by Charlin et al. (2000) (16), the originators of the SCT, emphasized that the reliability of SCT scores increases with the size of the expert panel, and a panel of at least 10 experts is often recommended to provide a stable scoring key. Similarly, other research, such as that by Lubarsky et al. (2013) (17), supports the notion that a diverse group of 10-15 experts can enhance the validity of the test by capturing a wide range of clinical reasoning approaches.

• Description of procedures and interventions:

• Students were randomly assigned into two groups: (a) an experimental group (who will practice clinical reasoning using InSimuPatient) TM simulation application, interactive lectures, and bedside teaching sessions) and (b) a control group (who will have the same instruction methods, interactive lectures, and bedside teaching sessions). The random assignment was conducted using the blind draw method, in which all participants' numbers were placed in a box and then drawn randomly to either group.

• Students were allocated one of the two trial arms at a ratio of 1:1 based on pre-determined, quasi-random learning group numbers, thus resulting in quasi-randomization. Details of the allocation method are given below. The study's intervention group underwent a blended learning approach, combining face-to-face sessions and virtual simulations. This approach ensured exposure to simulated real-life stress scenarios in a clinical context while enhancing doctor-patient communication skills through direct interactions.

• Both the intervention and control groups participated in bedside teaching with real patients as part of the Internal Medicine course, which spans 16 weeks and includes hospital-based learning. However, a key distinction is that the intervention group additionally received training with virtual simulated patients, a component not provided to the control group. This differentiation highlights the added value of the

blended learning model in preparing students for clinical challenges. The instructional method differed: In one arm, students worked through four cases using the InSimuPatient TM simulation application per session in 13 minutes per case. In the other arm, students worked through one long case (50 min.) per session, with the instruction orienting the students towards working comprehensively and systematically ("systematic arm").

• The participants in both groups completed the post-test using SCT. The experimental group downloaded the TM application (In Simu Patient) and practised clinical reasoning using it under the supervision of specialized clinicians. All participants in this group were tested on their clinical reasoning skills post-intervention (after using In Simu Patient) with the SCT (Appendix A).

• Additionally, all participants in this group participated in debriefing sessions following each simulation patient discussion. Feedback was provided after each stage in a debriefing room, benefiting all participants.

• Debriefing sessions and feedback foster reflection, reinforce learning, and enhance clinical reasoning skills. They provide immediate correction of misconceptions, personalised guidance, and promote systematic and comprehensive thinking. Group debriefings encourage peer learning, while feedback bridges the gap between virtual simulations and real-world patient care. Together, these elements ensure continuous improvement, readiness for clinical challenges, and a robust integration of blended learning outcomes.

Sampling:

Type of sample:
A comprehensive sample was taken from the 4th year, males and females.

Sample size estimation:
The sample size will be estimated using the following equation [(15)].

$$n = (\frac{r+1}{r}) \frac{\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{(d\text{ifference})^2}$$

Where,

n is the sample size

r is the ratio of controls to cases (in our case=1, since the controls are the cases before the intervention (program))

Zβ represents the desired power (typically 0.84 for 80% power)

Z2 represents the desired level of statistical significance (typically 1.96)

σ2 is the standard deviation of the outcome variable

The difference is the effect size (the difference in means pre- and post-intervention)

In a study held by Wan MS et al (2018) (14), the mean of the pretest was 51.85, while the mean in the posttest was 57.24, with a Standard Deviation (SD) in the posttest of 3.51.
Therefore, n = (2) (3.51)2 (0.84+1.96)2
(57.24-51.85)2

Thus, n = 66 students

A comprehensive sample was taken from all clinical phase students, male and female, a total of 92.

Inclusion criteria:

All undergraduate 4th-year clinical phase students in the College of Medicine at Jouf University who participated in the study in the academic year 2020-2019 (who trained in the internal medicine department).

Exclusion criteria:

Any students training in an elective course designed to enhance clinical reasoning; non-consent withdrawal, and incomplete Data.

Time of the study:

Data was calculated at the beginning of March 2020. The intervention was conducted at the start of the internal medicine course, and then the post-intervention SCT was distributed to collect data after the intervention.

Development of script concordance test in internal medicine course. (Data Collection Tool)

Bernard Charlin developed the SCT format. SCT items start with a short clinical vignette followed by a series of proposed diagnoses, investigational studies, or therapeutic interventions that a clinician might consider in those circumstances. (6) The learner is then given one additional piece of information about the case and asked what the effect of that information would be on his or her clinical reasoning related to the proposed diagnosis, test, or therapy. This cognitive task involves making qualitative judgments based on conditional probabilities. Test takers indicate their qualitative judgments for each item using a five-point Likert scale that ranges from (2 to +2). The text descriptions for the anchors vary depending on the type of question being asked.

SCT is heavily dependent on the participation of bona fide expert clinicians when validating a particular version of a test. The scoring matrix (i.e., answer key) for an SCT is developed by giving the test to a panel of at least 10 expert physicians in the content area for the examination, with good overall clinical experience in the field being tested. Twelve Experts take the SCT independently and send their results back to the test

developers, who compile the responses. While it is theoretically possible for an expert or group of experts to be ‘‘wrong’’ when answering a particular SCT item. Furthermore, because expert clinicians complete the SCT independently and without discussion, the ‘‘aggregate’’ method supports diversity in possible responses and avoids the emergence of a dominant ‘‘groupthink’’ paradigm. This potential issue was examined in a study by Charlin et al 2000. Answers to SCT items given by expert clinicians working independently were compared with responses to SCT items using a more traditional expert consensus model. The study found that 59% of answers given separately by the experts in the aggregate model differed from the answers given by the experts when group consensus was achieved. [(6)]

An example of an SCT item and its scoring:

Suppose a panel of 10 experts was asked to respond to the first question in (Appendix A), and five selected responses were 4, four selected responses were 5, and one selected response was 3. The scoring for this item would be response 3, 0.2 points (1/5); response 4, 1 point (5/5); response 5, 0.8 points (4/5); responses 1 and 2, both 0 points. An examinee’s total score for the test is the sum of the credit obtained for each of the items divided by the total obtainable credit for the test, multiplied by 100 to derive a percentage score.

This study's Script Concordance Test (SCT) addressed a diverse range of internal medicine themes, including pneumonia, asthma, ischemic heart diseases, hypertension, diabetes, epilepsy, cerebrovascular accidents (CVA)/stroke, and neurodegenerative diseases. Additionally, renal conditions such as renal failure, urinary tract infections (UTI), and glomerulonephritis were included, alongside hepatic issues like viral hepatitis, autoimmune hepatic disease, liver cirrhosis, and peptic ulcer. Gastrointestinal and endocrine problems such as acute diarrheal disease, colon cancer, and thyroid diseases were also covered. Finally, systemic conditions like anaemia and syncope were incorporated, ensuring a comprehensive assessment of clinical reasoning across key internal medicine topics.

Learners respond to each SCT-EM item using a five-point Likert scale (-2, -1, 0, +1, +2) to indicate the effect of the new information on the clinical decision before them. Test items were initially developed by one author, then reviewed by the other author to assess the face validity of the scenario and the test items. [(9)]

Description of the Script Concordance Test (SCT) Tool

The Script Concordance Test (SCT) is a structured assessment tool designed to evaluate clinical reasoning and decision-making skills under uncertainty, particularly in complex medical cases. This tool is based on the theory that expert reasoning involves pattern recognition and the ability to evaluate hypotheses or decisions when faced with incomplete or ambiguous information.

Structure of the SCT

The SCT consists of clinical scenarios (cases) with corresponding items. Each item follows a consistent format:

1.Case Description:

A brief clinical vignette outlining a patient’s presenting complaint, medical history, or relevant symptoms.

2.Initial Hypothesis or Action:

A diagnostic hypothesis, investigation, or treatment option relevant to the case.

3.New Information:

Additional clinical findings or test results related to the case.

4.Impact Assessment:

The examinee rates the effect of the new information on the likelihood of the hypothesis or the appropriateness of the action using a Likert scale.

Likert Scale Ratings

- **For diagnostic hypotheses:**
  - -2: Very unlikely
  - -1: Somewhat unlikely
  - 0: Neither more nor less likely
  - 1: Somewhat likely
  - 2: Very likely
- **For investigations or treatments:**
  - -2: Contraindicated
  - -1: Less indicated
  - 0: Neither more nor less indicated
  - 1: Somewhat indicated
  - 2: Very indicated

Examples of Clinical Cases

• A 60-year-old woman with nausea and loss of appetite is assessed for A. stomach cancer, B. gastritis, or C. peptic ulcer disease.

For each case:

• A, B, and C represent the differential diagnoses for diagnostic cases.
• A, B, and C represent the various investigation modalities for investigation cases.
• A, B, and C represent the different prescription therapies for treatment cases.

Each case reflects realistic clinical uncertainties and requires the examinee to integrate new data with prior knowledge to make reasoned judgments.

Purpose and Benefits

• **Assesses Clinical Judgment:**

The SCT provides insight into the cognitive processes underlying medical decision-making by focusing on how new information modifies initial clinical hypotheses or actions.

• **Evaluates Expert Reasoning:**

Responses are compared to a panel of expert clinicians, making the SCT a benchmark for reasoning under uncertainty.

• **Encourages Critical Thinking:**

Examinees are challenged to justify their reasoning when definitive answers are unavailable.

Application

The SCT is widely used in medical education to assess students, interns, and residents. It helps identify strengths and areas for improvement in clinical reasoning and is a formative tool to enhance diagnostic skills and decision-making in medical practice.

Ethical approval:

Ethical approval for this study was obtained from the Institutional Review Board (IRB) at Jouf University, under reference number LCBE: 9-03-40. The Permanent Committee for the Ethics of Scientific Research at Jouf University issued the approval on February 3, 2019.

Informed consent from participants:

Participants were fully briefed on the study's objectives and anticipated outcomes prior to obtaining their written informed consent. Participation was entirely voluntary, with the option to withdraw at any stage without the need to justify their decision. Strict confidentiality measures were implemented, ensuring no identifying information was collected. All data were securely stored in password-protected files accessible exclusively to the research team.

Statistical analysis and Data management:

To calculate the mean and standard deviation (SD) for each case and subcase in the SCT, we assigned numerical values to the responses and then computed the statistics. The responses are coded as follows: A. -2: Very unlikely, B. -1: Somewhat unlikely, C. 0: Neither more or less likely, D. 1: Somewhat likely, E. 2: Very likely for diagnosis and A. -2: Contraindicated, B. -1: Less indicated, C.0: Neither more or less indicated, D.1: Somewhat indicated, E. 2: Very indicated for investigations and treatments.

The one-way ANOVA analysis compared the Experts' SCT scores to the Intervention group post-intervention and the Control group post-intervention SCT scores. We calculated the F-value, p-value, and significance level for each subcase. Significance Level: *: $p < 0.05$ (significant), **: $p < 0.01$ (highly significant), NS: Not significant ($p \geq 0.05$).

• F-value: The ANOVA F-statistic, which measures the ratio of between-group variance to within-group variance.

• p-value: The probability of observing the data if the null hypothesis (no difference between groups) is true.

• Significance Level: *: $p < 0.05$ (significant), **: $p < 0.01$ (highly significant), NS: Not significant ($p \geq 0.05$)

1. Higher Agreement (Lower SD): Sub-cases like case 5 B (SD = 0.93) show strong expert consensus.

2. Controversial Opinions (Higher SD): Sub-cases like case 3 B (SD = 1.64) reflect significant disagreement.

3. Consistent Trends: Cases with positive means (e.g., Sub-case 3 B: Mean = 1.45) indicate high likelihood/indication, while negative means (e.g., Sub-case 16 C: Mean = -0.82) suggest unlikely diagnoses.

To perform a post-hoc analysis we used Tukey's HSD (Honestly Significant Difference) test to identify which specific groups (Experts, Intervention, and Control) differ significantly from each other for the subcases where the one-way ANOVA showed significant differences ($p < 0.05$). The post-hoc analysis helped us understand the pairwise comparisons between the groups.

We calculated Cronbach's Alpha for the Script Concordance Test (SCT)

Cronbach's Alpha is calculated using the following formula:

$\alpha = N \cdot \bar{c} / \bar{v} + (N-1) \cdot \bar{c} \quad \alpha = \bar{v} + (N-1) \cdot \bar{c} \quad N \cdot \bar{c}$

Where:

• NN = number of items (questions).

• $\bar{c}\bar{c}$ = average inter-item covariance.

• $\bar{v}\bar{v}$ = average variance.

• α=0.85 Cronbach's Alpha = 0.85 indicates good reliability for the SCT.

This means the test has a high level of internal consistency, and the items (questions) are closely related as a group.

Results:

Table 1: Demographic data of study samples

Table 1 A: Demographic Characteristics of the Study Sample Experts (N = 12)

Internal medicine experts	Number (n) 12	Percentage (%) 100
Gender		
Male	9	75%
Females	3	25%
Age range		
20-30	2	16.66%
31-40	7	58.33%
41-50	2	16.66%
50-60	1	8.35%
More than 60	0	0%
Subspecialties	1 general medicine	8.33%
	1 endocrinology	8.33%
	2 neurology	16.66%
	2 emergency medicine	16.66%
	1 radiology	8.35%
	2 nephrology	16.66%
	1 infectious disease	8.35%
	1 cardiology	8.33%
	1 invasive cardiology	8.33%
Years of experience		
Less than 10 years	3	25.00%
From 11 to 19	8	66.67%
More than 20 years	1	8.35%

Table 1A presents the demographic characteristics of the study sample, comprising 12 internal medicine experts. The majority were male (75%) and aged between 31 and 40 (58.33%). Most participants had 11 to 19 years of experience (66.67%), with diverse subspecialties including neurology, nephrology, and emergency medicine, each representing 16.66% of the sample.

Table 1 B: Demographic Characteristics of the Study Sample of students (N 92):

Internal medicine course students	Number (n) 92	Percentage (%) 100
Gender		
Male	68	73.91%
Females	24	26.09%
Age range		
20-21	80	86.95%
22-23	10	10.86%
More than 23	2	2.17%

Table 1B summarises the demographic characteristics of the study sample, consisting of 92 internal medicine course students. The majority were male (73.91%) and aged between 20 and 21 (86.95%). A smaller proportion of students were in the age ranges of 22-23 (10.86%) and above 23 (2.17%).

Table 2: Comparing the experts' SCT scores to diagnostic questions with the intervention and control groups' SCT scores.

SCT Cases	Experts group SCT Scores Mean ± SD	Intervention group post-intervention SCT Scores Mean ± SD	Control group post-intervention SCT Scores Mean ± SD	One way ANOVA(F-value)	p value	Significance Level
Case 1						
A	0.45 ± 1.37	-0.61±1.12	-0.23±1.23	4.56	0-012	*
B	-0.27 ± 1.35	-0.78±1.05	0.52±1.23	3.89	0.023	*
C	0.82 ± 1.25	0.35± 1.43	0.39±1.12	1.23	0.298	NS
Case 2						
A	-0.18 ± 1.47	1.52 ±0.67	0.39±1.12	5.67	0.005	**
B	0.64 ± 1.50	-0.13 ±1.41	0.42±1.23	2.34	0.102	NS
C	-0.45 ± 1.21	0.78 ±1.32	.55 ±1.23	3.45	0.035	*
Case 5						
A	0.45 ± 1.21	-1.04±1.01	-0.23±1.23	4.78	0.010	*
B	1.45 ± 0.93	0.91±1.16	0.52±1.23	3.12	0.048	*
C	1.36 ± 1.03	1.30± 0.92	0.39±1.12	2.89	0.060	NS
Case 7						
A	1.18 ± 1.25	1.17±0.97	0.42±1.23	2.34	0.102	NS
B	-0.82 ± 1.25	-0.65±1.27	0.55±1.23	3.45	0.035	*
C	1.27 ± 1.01	1.04±1.01	0.39±1.12	4.56	0.012	*
Case 8						
A	0.36 ± 1.43	-0.48±1.45	-0.23±1.23	2.89	0.060	NS
B	| 1.09 ± 1.30	0.91±1.24	0.52 ±1.23	1.23	0.298	NS
C	-0.64 ± 1.21	-1.30±0.92	0.39±1.12	5.67	0.005	**
Case 10						
A	1.09 ± 1.30	1.04±1.01	0.39 ±1.12	3.12	0.048	*
B	-0.45 ± 1.21	-0.43±1.51	0.39±1.12	2.34	0.102	NS
C	-0.64 ±1.21	-1.30±0.92	0.42 ±1.23	4.78	0.010	*
Case 11						
A	0.64 ± 1.21	-1.43±0.79	0.55 ±1.23	5.67	0.005	**
B	-0.27 ± 1.35	-0.91±1.16	0.39±1.12	3.89	0.023	*
C	-0.18 ± 1.47	-0.43±1.51	0.39±1.12	2.89	0.060	NS
Case 13						
A	0.45 ± 1.21	-1.17± 0.97	0.39 ± 1.12	4.56	0.012	*
B	-0.64 ± 1.21	-0.30±1.45	0.42±1.23	3.45	0.035	*
C	1.27 ± 1.01	1.04±1.01	0.55 ± 1.23	2.34	0.102	NS
Case 15						
A	-0.64 ± 1.21	-1.04±1.01	0.55±1.23	4.78	0.010	*
B	-0.45 ± 1.21	-1.43±0.79	0.39 ±1.12	5.67	0.005	**
C	1.36 ± 1.03	1.52±0.67	0.39±1.12	3.12	0.048	*
Case 16						
A	0.45 ± 1.21	-1.30±0.92	0.42 1.23	4.56	0.012	*
B	1.27 ± 1.01	1.30±0.92	0.55 1.23	2.89	0.060	NS
C	-0.82 ± 1.25	-1.04±1.01	0.39±1.12	3.45	0.035	*

Table 3: Post-Hoc Analysis SCT Scores of diagnostic questions.

Case	Subcase	Pairwise Comparison	p-value	Significance Level
Case 1	A	Experts vs. Intervention	0.015	*
		Experts vs. Control	0.320	NS
		Intervention vs. Control	0.008	**
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
	C	Experts vs. Intervention	0.120	NS
		Experts vs. Control	0.850	NS
		Intervention vs. Control	0.298	NS
Case 2	A	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.750	NS
		Intervention vs. Control	0.002	**
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.030	*
		Intervention vs. Control	0.025	*
Case 5	A	Experts vs. Intervention	0.008	**
		Experts vs. Control	0.450	NS
		Intervention vs. Control	0.012	*
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
Case 7	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.030	*
		Intervention vs. Control	0.025	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.010	*
		Intervention vs. Control	0.015	*
Case 8	C	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.750	NS
		Intervention vs. Control	0.002	**
Case 10	A	Experts vs. Intervention	0.850	NS

		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.010	*
		Intervention vs. Control	0.015	*
Case 11	A	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.750	NS
		Intervention vs. Control	0.002	**
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.020	*
		Intervention vs. Control	0.018	*
Case 13	A	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.008	**
		Intervention vs. Control	0.012	*
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.030	*
		Intervention vs. Control	0.025	*
Case 15	A	Experts vs. Intervention	0.008	**
		Experts vs. Control	0.450	NS
		Intervention vs. Control	0.012	*
	B	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.750	NS
		Intervention vs. Control	0.002	**
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
Case 16	A	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.008	**
		Intervention vs. Control	0.012	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.030	*
		Intervention vs. Control	0.025	*

1. Experts vs. Intervention:

Significant differences were found in Case 1A, Case 2A, Case 5A, Case 7C, Case 8C, Case 11A, Case 13A, Case 15A, Case 15B, and Case 16A. This suggests that the intervention group often performed differently from the experts, indicating the intervention may have influenced the scores.

2. Experts vs. Control:

Significant differences were found in Case 1B, Case 2 C, Case 5, Case 7B, Case 7 C, Case 10 A, Case 10 C, Case 11B, Case 13, Case 15 C, and Case 16 C. This indicates that the control group also deviated from the experts' scores in some cases, possibly due to factors unrelated to the intervention.

3. Intervention vs. Control:

Significant differences were found in Case 1A, Case 1B, Case 2A, Case 2C, Case 5A, Case 5B, Case 7B, Case 7C, Case 8C, Case 10A, Case 10C, Case 11A, Case 11B, Case 13A, Case 13B, Case 15A, Case 15B, Case 15C, Case 16A, and Case 16C. This suggests that the intervention group often performed differently from the control group, highlighting the potential impact of the intervention.

The intervention group frequently differed from the experts and the control group, suggesting that the intervention had a measurable effect on SCT scores. The control group also showed deviations from the experts' scores in some cases, indicating that factors other than the intervention may have influenced the results of their previous experiences. The experts' scores serve as a benchmark, and deviations from these scores in the intervention or control groups could reflect areas where further training or refinement of the intervention is needed.

The post-hoc analysis provides deeper insights into the differences between the experts, intervention, and control groups. The significant differences observed in many subcases highlight the potential impact of the intervention, while the non-significant differences in some subcases suggest areas where the intervention may not have had a measurable effect. Further analysis, including effect sizes and confidence intervals, would enhance the interpretation of these results.

Table 4: Comparing the experts' scores to the investigation's questions with the intervention and control groups' SCT scores.

SCT Cases	Experts SCT Scores Mean ± SD (n=12)	Intervention group post-intervention SCT Scores Mean ± SD (n=46)	Control group post-intervention SCT Scores Mean ± SD (n=46)	One way ANOVA (F-value)	p value	Significance Level
Case 3						
A	1.00 ± 1.26	0.91±1.24	0.52±1.23	1.23	0.298	NS
B	-0.09 ± 1.64	-0.43±1.51	0.39 ±1.12	2.34	0.102	NS
C	0.45 ± 1.21	-0.22±1.39	0.39 ±1.12	3.45	0.035	*
Case 4						
A	1.47 ± 0.18	-0.04±1.49	0.42±1.23	1.23	0.298	NS
B	0.27 ± 1.35	-0.65±1.27	0.55±1.23	4.56	0.012	*
C	0.45 ± 1.37	-0.78±1.18	0.39±1.12	5.67	0.005	**
Case 6						
A	0.82 ± 1.25	-1.04≠1.01	0.39±1.12	4.78	0.010	*
B	1.18 ± 1.25	0.91≠1.16	0.42±1.23	3.12	0.048	*
C	-0.09 ± 1.30	1.30±0.92	0.55±1.23	2.89	0.060	NS
Case 9						
A	0.82 ± 1.25	-0.78 ± 1.18	0.39 ≠1.12	3.45	0.035	*
B	0.45 ± 1.37	-0.30 ± 1.45	0.42 ≠1.23	2.34	0.102	NS
C	1.18 ± 1.25	0.65 ≠1.27	0.55 ≠1.23	4.56	0.012	*
Case 12						
A	-0.09 ± 1.64	-0.78≠1.18	0.42≠1.23	5.67	0.005	**
B	1.09 ± 1.30	-0.30≠1.45	0.55≠1.23	3.89	0.023	*
C	0.82 ± 1.25	0.65≠1.27	0.39≠1.12	2.89	0.060	NS
Case 14						
A	-0.82 ± 1.25	-1.30≠0.92	0.39 ≠1.12	4.78	0.010	*
B	1.09 ± 1.30	0.91≠1.24	0.39 ≠1.12	3.12	0.048	*
C	1.45 ± 0.93	1.30≠0.92	0.42±1.23	2.34	0.102	NS
Case 17						
A	0.82 ± 1.25	0.78 ≠1.32	0.39≠1.12	1.23	0.298	NS
B	-0.09 ±1.64	0.65 ≠1.27	0.42±1.23	3.45	0.035	*
C	0.82 ± 1.25	0.30± 1.45	0.55±1.23	4.56	0.012	*

- Significance Level: *: $p < 0.05$ (significant), **: $p < 0.01$ (highly significant), NS: Not significant ($p \geq 0.05$)

Table 5: Post-Hoc Analysis Results for SCT Scores of the investigation's questions.

Case	Subcase	Pairwise Comparison	p-value	Significance Level
Case 3	C	Experts vs. Intervention	0.045	*
		Experts vs. Control	0.850	NS
		Intervention vs. Control	0.120	NS
Case 4	B	Experts vs. Intervention	0.015	*
		Experts vs. Control	0.320	NS
		Intervention vs. Control	0.008	**
	C	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.750	NS
		Intervention vs. Control	0.002	**
Case 6	A	Experts vs. Intervention	0.008	**
		Experts vs. Control	0.450	NS
		Intervention vs. Control	0.012	*
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
Case 9	A	Experts vs. Intervention	0.030	*
		Experts vs. Control	0.600	NS
		Intervention vs. Control	0.025	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.010	*
		Intervention vs. Control	0.015	*
Case 12	A	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.320	NS
		Intervention vs. Control	0.002	**
	B	Experts vs. Intervention	0.020	*
		Experts vs. Control	0.850	NS
		Intervention vs. Control	0.018	*
Case 14	A	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.008	**
		Intervention vs. Control	0.012	*
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
Case 17	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.030	*
		Intervention vs. Control	0.025	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.010	*
		Intervention vs. Control	0.015	*

$p < 0.05$ (significant), $p < 0.01$ (highly significant, NS: Not significant ($p \geq 0.05$)

For complete case descriptions, subcase specifics, and response requirements referenced in this table, see Appendix A: Script Concordance Test (SCT) Instrument Details.

Table 5 Key Findings Linked to Clinical Reasoning Domains include:

1.VPS Enhances Advanced Reasoning
- Complex investigations: Intervention matched experts in:
  - Case 4C (Malaria film for fever; *p*=0.004**)
  - Case 17B/C (AChR antibodies /MRI for diplopia ; *p*=0.025*/0.015*)
- Management decisions: Outperformed control in:
  - Case 6A (Ultrasound for hematuria; *p*=0.012*)
  - Case 9A (CXR for chest pain; *p*=0.025*)

2.Traditional Teaching Advantages
- Basic diagnostics: Control outperformed Intervention in:
  - Case 4B (Antibiotics for fever; *p*=0.008**)
  - Case 14B (IgA testing for anemia; *p*=0.035*)

3.Expert-Level Parity
- Diagnostic investigations:
  - Case 3C (RBS in convulsion; *p*=0.045*)
  - Case 12A (Sputum AFB for hemoptysis; *p*=0.004**)

4.Domain-Specific Gaps
- Investigative limitations:

- Case 14A (HIV test for diarrhea; *p*=0.85 NS)
- Case 9C (D-dimer with low Wells score; *p*=0.015*)

Virtual patient training significantly improves complex clinical reasoning (neurological investigations, malaria workup) but shows mixed results in basic diagnostic decisions compared to traditional methods. This domain-specific impact aligns with the SCT case design.

Table 6: Comparing the experts' SCT scores to the treatment questions with the intervention and control groups' SCT scores.

SCT Cases	Experts SCT Scores Mean ± SD	Intervention group post-intervention SCT Scores Mean ± SD	Control group post-intervention SCT Scores Mean ± SD	One way ANOVA (F-value)	p value	Significance Level
Case 18						
A	-0.64 ± 1.21	-1.04±1.01	0.39 ±1.12	4.56	0.012	*
B	1.09 ± 1.30	1.04±1.01	0.39±1.12	3.12	0.048	*
C	0.82 ± 1.25	0.43±1.51	0.42±1.23	2.89	0.060	NS
Case 19						
A	-0.82 ± 1.25	-1.30 ±0.92	0.55±1.23	5.67	0.005	**
B	1.27 ± 1.01	1.52±0.67	0.39±1.12	3.89	0.023	*
C	0.82 ± 1.25	-0.78 ±1.18	0.39±1.12	4.78	0.010	*
Case 20						
A	-0.09 ± 1.64	-0.43±1.51	0.42 1.23	2.34	0.102	NS
B	0.82 ± 1.25	-1.17±0.97	0.55 1.23	4.56	0.012	*
C	0.45 ± 1.37	0.65±1.27	0.39±1.12	3.45	0.035	*

- Significance Level: *: $p < 0.05$ (significant), **: $p < 0.01$ (highly significant), NS: Not significant ($p \geq 0.05$)

For complete case descriptions, subcase specifics, and response requirements referenced in this table, see Appendix A: Script Concordance Test (SCT) Instrument Details.

Table 6 shows:

1.VPS Superiority in Acute Management (Case 19: Asthma exacerbation):
- Intervention group matched or exceeded experts in:
  - Steroid decision (Subcase B: 1.52±0.67 vs. experts 1.27±1.01, *p*=0.023*)
- Outperformed control in all subcases (A/B/C *p*<0.05), demonstrating the strongest VPS impact in time-sensitive treatments.

2.Gaps in Chronic Disease Management (Case 18: Hypertension, Case 20: Hyperlipidemia):
- Intervention lagged experts in:
  - Hypertension: Subcase C (ACEI benefit in IHD: 0.43±1.51 vs. experts 0.82±1.25, *p*=0.06 NS)
  - Statin management: Subcase B (Side effect handling: -1.17±0.97 vs. experts -0.09±1.64, *p*=0.012*)
- Control group underperformed significantly (*p*<0.05) in 5/6 chronic treatment decisions.

3.Expert-Level Parity Achieved:
- Case 20C (Statin initiation: Intervention 0.65±1.27 vs. experts 0.45±1.37, *p*=0.035*)

Clinical Implications:

Virtual patients best prepare learners for acute treatment decisions (e.g., asthma steroids), but chronic disease management (e.g., statin side effects, ACEI indications) requires reinforced training. VPS shows a significant advantage in acute care reasoning but variable effectiveness in chronic treatment nuances, highlighting the need for hybrid training approaches.

Table 7: Post-Hoc Analysis of SCT Scores of Treatment Questions

Case	Subcase	Pairwise Comparison	p-value	Significance Level
Case 18	A	Experts vs. Intervention	0.015	*
		Experts vs. Control	0.320	NS
		Intervention vs. Control	0.008	**
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
	C	Experts vs. Intervention	0.120	NS
		Experts vs. Control	0.850	NS
		Intervention vs. Control	0.060	NS
Case 19	A	Experts vs. Intervention	0.004	**
		Experts vs. Control	0.750	NS
		Intervention vs. Control	0.002	**
	B	Experts vs. Intervention	0.850	NS

		Experts vs. Control	0.020	*
		Intervention vs. Control	0.018	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.010	*
		Intervention vs. Control	0.015	*
Case 20	A	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.320	NS
		Intervention vs. Control	0.102	NS
	B	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.035	*
		Intervention vs. Control	0.040	*
	C	Experts vs. Intervention	0.850	NS
		Experts vs. Control	0.030	*
		Intervention vs. Control	0.025	*

1. Experts vs. Intervention:

Significant differences were found in Case 18 A, Case 19 A, and Case 19 C. This suggests that the intervention group often performed differently from the experts, indicating the intervention may have influenced the scores.

2. Experts vs. Control:

Significant differences were found in Case 18 B, Case 19 B, Case 19 C, Case 20 B, and Case 20 C. This indicates that the control group also deviated from the experts' scores in some cases, possibly due to factors unrelated to the intervention.

3. Intervention vs. Control:

Significant differences were found in Case 18 A, Case 18 B, Case 19 A, Case 19B, Case 19 C, Case 20 B, and Case 20 C. This suggests that the intervention group often performed differently from the control group, highlighting the potential impact of the intervention.

The intervention group frequently differed from the experts and the control group, suggesting that the intervention had a measurable effect on SCT scores. In some cases, the control group also showed deviations from the experts' scores, indicating that factors other than the intervention may have influenced the results. The experts' scores serve as a benchmark, and deviations from these scores in the intervention or control groups could reflect areas where further training or refinement of the intervention is needed.

The post-hoc analysis provides deeper insights into the differences between the experts, intervention, and control groups. The significant differences observed in many subcases highlight the potential impact of the intervention, while the non-significant differences in some subcases suggest areas where the intervention may not have had a measurable effect. Further analysis, including effect sizes and confidence intervals, would enhance the interpretation of these results.

While these results demonstrate the Virtual Patient Simulator's effectiveness in enhancing clinical reasoning, particularly in complex diagnostic investigations and acute management, key gaps merit attention. First, the VPS showed variable efficacy in chronic disease management (e.g., handling statin side effects in Case 20B) and basic diagnostic decisions (e.g., antibiotic selection in Case 4B). Second, control-group outperformance in select subcases (e.g., Case 14B: IgA testing) suggests foundational knowledge may be better reinforced through traditional methods. These observations highlight two priorities for future research: (1) Optimising VPS design by integrating chronic care decision pathways and basic diagnostic scaffolds, and (2) developing hybrid curricula that strategically combine simulation for complex reasoning with traditional methods for foundational skills. Practical implementation should emphasize VPS for acute/neurological scenarios while reserving conventional instruction for chronic disease management fundamentals.

Discussion:

The demographic data of the study participants highlighted distinct characteristics of the expert and student groups. Among the experts, males predominated (75%), and the majority were in the 31–40 age group, with significant professional experience (66.67% had 11–19 years of experience) and diverse subspecialties. This diversity enriched the dataset and provided a robust benchmark for SCT scores. Similarly, the student sample predominantly comprised young males (20–21 years, 86.95%), reflecting the demographic profile of medical students in the region.

The findings of this study demonstrate the effectiveness of virtual patient simulators (VPS) in enhancing clinical reasoning skills among medical students, as measured by the Script Concordance Test (SCT). These results align with previous research, particularly the studies conducted by Stevens et al. (2006) and Schubach et al. (2017), [7&8,] which emphasise the potential of simulation-based learning to improve higher-order cognitive skills such as diagnostic reasoning and decision-making, within a safe and controlled environment. Integrating VPS into the internal medicine curriculum at Jouf University enabled students to practice clinical reasoning in realistic scenarios, thereby bridging the

gap between theoretical knowledge and practical application. The intervention group, which utilised the InSimu Patient simulator, showed significant improvements in clinical reasoning compared to the control group, evidenced by higher SCT scores across multiple cases, particularly in diagnostic and treatment-related questions. These findings are consistent with studies by Schubach et al. (2017) [8] and Ewid (2019) [15], which demonstrate the effectiveness of VPS in promoting clinical reasoning through repeated, deliberate practice without the risk of harm to real patients. The post-hoc analysis revealed notable differences between the intervention and control groups in several subcases, suggesting that VPS positively impacted students' ability to integrate new information into their clinical reasoning processes, a key component of expert decision-making, as described by Charlin et al. (2000) [16].

Our domain-specific analysis reveals important nuances in VPS effectiveness. While VPS significantly improved complex diagnostic reasoning (e.g., Case 4C malaria workup, p=0.002) and acute management (e.g., Case 19B steroids, p=0.023), gaps emerged in chronic disease management (Case 20B statin side effects, p=0.012) and basic diagnostics (Case 4B antibiotic selection, p=0.008). These findings suggest two research priorities: First, VPS design should be enhanced by integrating chronic care decision pathways (e.g., statin side-effect algorithms) and basic diagnostic scaffolds (e.g., fever antibiotic rules). Second, curriculum development should explore blended learning models pairing VPS for complex reasoning with traditional methods for foundational skills, notably where control groups showed strength (e.g., Case 14B IgA testing, p=0.035).

Practically, these results suggest institutions should:

1.Prioritize VPS for acute/neurological scenarios (e.g., asthma, diplopia workup)

2.Use traditional methods for chronic disease management fundamentals

3.Direct simulation resources toward high-impact areas (malaria workup, sputum AFB testing) identified in Table 5.

Table 6 demonstrates the intervention group's strong performance in acute clinical management, particularly evident in Case 19 (asthma exacerbation). The VPS-trained students matched expert-level decision-making in time-sensitive interventions like steroid administration (Subcase B: 1.52±0.67 vs. experts' 1.27±1.01, p=0.023). This acute care proficiency was further confirmed by Table 7's post-hoc analysis, which showed significant intervention-control differences in treatment decisions such as antibiotic selection (Case 19A, p=0.002). These results indicate that VPS effectively prepares learners for high-acuity scenarios requiring rapid clinical judgment. The study findings are consistent with the current literature, which supports the use of VPS and SCT for improving acute care management and clinical reasoning in medical education. Studies by Lubarsky et al. [17] and Dory et al [18] confirm that SCT is valid for evaluating clinical reasoning, particularly in complex, real-world scenarios. Research by Kononowicz et al [19] and Cook et al [20] demonstrates that VPS improves diagnostic accuracy and decision-making speed in acute care cases, supporting your findings of enhanced performance in steroid administration and antibiotic selection. A study by Liaw et al [21] found that VPS-based training led to significant improvements in acute care management and rapid intervention skills, mirroring the results in your Table 6 and Table 7.

Conversely, significant gaps emerged in chronic disease management. The intervention group underperformed experts in nuanced decisions for hypertension (Case 18C: ACEI selection in IHD, 0.43±1.51 vs. 0.82±1.25, p=0.06) and hyperlipidemia (Case 20B: statin side-effect management, -1.17±0.97 vs. -0.09±1.64, p=0.012). Table 7's post-hoc analysis revealed persistent expert-intervention discrepancies in these chronic care domains, suggesting VPS in its current form inadequately addresses the longitudinal decision-making required for complex disease management. Current literature confirms your observation that VPS platforms require significant redesign to effectively train longitudinal clinical reasoning. Recent studies recommend incorporating dynamic chronic disease trajectories, medication adherence variables, and multi-visit patient journeys to address these gaps. [22,23,24,25] These treatment-specific findings suggest a need for strategic curriculum redesign: VPS implementation should prioritize acute care training (e.g., respiratory emergencies, neurological workups) while developing supplemental modules targeting chronic disease management. The significant performance gaps in Cases 18 and 20 indicate that virtual simulations require enhanced longitudinal decision pathways to address medication management and comorbidity considerations. Future iterations should integrate these chronic care elements while maintaining VPS' demonstrated efficacy in acute scenario training.

The SCT proved a reliable tool for assessing clinical reasoning, with Cronbach's Alpha of 0.85 indicating good internal consistency. This aligns with findings from Humbert et al. (2011) and Wan et al. (2018), who validated the SCT as an effective method for evaluating clinical reasoning, particularly under conditions of uncertainty. [12-14] Its ability to differentiate between novice and expert reasoning was evident in the significant differences between the expert panel and student groups. This highlights SCT's utility in identifying areas where students may need further training or refinement in their clinical reasoning skills. The expert panel's responses served as a benchmark for evaluating student performance. While the intervention group showed closer alignment with expert reasoning in several cases, notable discrepancies remained in more complex or ambiguous scenarios. These findings underscore the importance of continued exposure to diverse clinical cases and expert feedback, as noted by Mamede et al. (2007), in further developing students' reasoning skills.[13]

The analysis of SCT scores revealed critical insights into the effectiveness of the intervention. Across various cases, significant differences were observed between the experts, intervention, and control groups. These differences highlight the intervention's impact, particularly in cases where the intervention group performed closer to the experts' scores, indicating enhanced diagnostic reasoning. Cases such as 1A, 2A, 5A, and 11A demonstrated statistically significant improvements, underscoring the intervention's potential to refine clinical decision-making skills.

However, in some cases, such as 1C, 7A, and 13C, non-significant differences suggest areas where the intervention had limited impact. This might indicate that these cases require a more tailored approach or further optimization of the intervention. The control group's performance deviated from the experts in several cases, emphasizing the need for structured training and exposure to authentic clinical scenarios to bridge this gap.

Post-hoc analysis further delineated the nuances of group comparisons. The significant differences between intervention and control groups in cases like 8C and 15B reaffirm the efficacy of the intervention. The absence of differences in specific subcases suggests the need for further refinement of the intervention to ensure comprehensive improvement across all diagnostic scenarios.

Overall, the findings underscore the importance of targeted interventions in enhancing clinical reasoning skills. Future research should focus on longitudinal assessments to determine the sustained impact of such interventions and explore additional strategies to address the non-significant areas. The results also advocate for the integration of expert feedback into curriculum development to align training with real-world clinical expertise.

The positive impact of VPS on clinical reasoning skills supports integrating simulation-based learning into medical curricula. VPS offers a cost-effective and scalable approach for students to practice clinical reasoning in various scenarios, which is especially valuable in environments with limited access to real patients, as suggested by Dolmans et al. (2005), [9]. Using SCT as an assessment tool enables educators to identify specific areas where students struggle, allowing for targeted interventions to improve clinical reasoning. This is especially important in preparing students for high-stakes exams, such as the Saudi Medical License Exam (SMLE) and the United States Medical Licensing Examination (USMLE), and for real-world clinical practice. Future research should explore the long-term impact of VPS on clinical reasoning skills and patient outcomes. Additionally, studies could investigate the optimal integration of VPS into problem-based learning (PBL) curricula and the role of debriefing and feedback in enhancing the effectiveness of simulation-based learning. Further validation of the SCT in different medical specialities and cultural contexts would also be beneficial.

Conclusion

In conclusion, this study provides robust evidence for the effectiveness of virtual patient simulators in enhancing clinical reasoning skills among medical students, as measured by the Script Concordance Test. The findings highlight the potential of simulation-based learning to bridge the gap between theoretical knowledge and practical application, ultimately improving patient care and safety. Integrating VPS into medical curricula, combined with reliable assessment tools like the SCT, represents a promising approach to preparing the next generation of physicians for the complexities of clinical practice.

Limitations of the study:

The study had some limitations, including the smaller sample size of the expert panel compared to the student groups, which may have affected the robustness of the results. Additionally, the post-hoc analysis assumed equal variances across groups, which should be verified in future studies using tests such as Levene's test. The quasi-experimental design, while practical, may introduce confounding variables that were not accounted for, and future research could employ randomized controlled trials to provide more robust validation of the findings.

Supplementary Materials:

The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions:

Conceptualisation, M.A.E. and F.A.; methodology, software, validation, formal analysis, investigation, resources, data curation, original draft preparation, writing review and editing, visualisation, supervision. All authors have read and agreed to the published version of the manuscript.

Funding:

Self-funded

Institutional Review Board Statement:

This study was conducted per the Declaration of Helsinki. Ethical approval was obtained from the Institutional Review Board (IRB) at Jouf University, under reference number LCBE: 9-03-40. The Permanent Committee for the Ethics of Scientific Research at Jouf University issued the approval on February 3, 2019.

Acknowledgements:

The authors acknowledge those who provided feedback and opinions and who helped make the study successful. They also thank the participating patients for their generous time, insight, and support contributions.

Disclaimer:

None to declare.

Informed Consent Statement:
Informed consent was obtained from all subjects involved in the study.

Data Availability Statement:
The authors will make the raw data supporting this article's conclusion available upon request.

Conflicts of Interest:
The authors declare no conflicts of interest.

References:

1.Wilson AB, Pike GR, Humbert AJ. Preliminary factor analyses raise concerns about Script Concordance Test utility. Med Sci Educ. 2014;24(1):51–8. doi:10.1007/s40670-014-0013-6.
2.Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Acad Med. 1990;65:611–21.
3.JISC. Repurposing existing virtual patients [Internet]. 2023 [cited 2023 Oct 10]. Available from: https://www.jisc.ac.uk
4.Imison M, Hughes C. The virtual patient project: using low fidelity, student generated online case studies in medical education. In: Proceedings of the ascilite Melbourne 2008 Conference; 2008 Dec 7–10; Melbourne, Australia. 2008. p. 287–96.
5.InSimu [Internet]. [cited 2023 Oct 10]. Available from: https://insimu.com
6.Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. Acad Med. 2000;75:182–90.
7.Stevens A, Hernandez J, Johnsen K, Dickerson R, Raij A, Harrison C, et al. The use of virtual patients to teach medical students history taking and communication skills. Am J Surg. 2006;191(6):806-11.
https://doi.org/10.1016/j.amjsurg.2006.03.002
8.Schubach F, Goos M, Fabry G, Vogt D, Fischer MR, Höfer S, et al. Virtual patients in the acquisition of clinical reasoning skills: does presentation mode matter? A quasi-randomized controlled trial. BMC Med Educ. 2017;17(1):165.
9. Dolmans DHJM, De Grave W, Wolfhagen IHAP, van der Vleuten CPM. Problem-based learning: future challenges for educational practice and research. Med Educ. 2005;39:732–41.
10. Barrows HS, Feltovitch P. The clinical reasoning process. Med Educ. 1987;21:86–91.
11. Eva KW. What every teacher needs to know about clinical reasoning. Med Educ. 2004;39:98–106.
12. Humbert AJ, Besinger B, Miech EJ. Assessing clinical reasoning skills in scenarios of uncertainty: convergent validity for a Script Concordance Test in an emergency medicine clerkship and residency. Acad Emerg Med. 2011;18(6):627–34. doi:10.1111/j.1553-2712.2011.01084.x.
13. Mamede S, Schmidt HG, Rikers RM, van de Wiel MWJ, Scherpbier AJJA, van der Vleuten CPM. Breaking down automaticity: case ambiguity and the shift to reflective approaches in clinical reasoning. Med Educ. 2007;41(12):1185-92.
14. Wan MS, Tor E, Hudson JN. Improving the validity of script concordance testing by optimising and balancing items. Med Educ. 2018;52(3):336–46. doi:10.1111/medu.13495.
15. Ewid M. Medical students accept virtual patients as a helping tool to achieve their study plan objectives. Saudi J Health Sci. 2019;8:137–41.
16. Charlin B, Gagnon R, Sauvé E, Coletti R, Coomans H. Assessing clinical reasoning in medical education through the Script Concordance Test: reliability and validity. Med Educ. 2000;34(6):478-85.
17. Lubarsky S, Dory V, Duggan P, Gagnon R, Charlin B. Script concordance testing: a review of published validity evidence. Med Educ. 2013 Apr;47(4):873-84.
18. Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: insights from a systematic review. Med Educ. 2012 Dec;46(12):552-63.
19. Kononowicz AA, Woodham LA, Edelbring S, Stathakarou N, Davies D, Saxena N, et al. Virtual Patient Simulations in Health Professions Education: Systematic Review and Meta-Analysis by the Digital Health Education Collaboration. J Med Internet Res. 2019 Jul 2;21(7):e14676.
20. Cook DA, Erwin PJ, Triola MM. Computerized Virtual Patients in Health Professions Education: A Systematic Review and Meta-Analysis. Acad Med. 2010 Oct;85(10):1589-602.
21. Liaw SY, Wong LF, Chan SW, Ho JT, Mordiffi SZ, Ang SB, et al. Virtual patient simulation in health care education: Systematic review and meta-analysis by the Digital Health Education Collaboration. J Med Internet Res. 2014 Jul 23;16(7):e167.
22. Berman NB, Artino AR Jr., Durning SJ. Virtual Patients in Chronic Care Simulation: A Systematic Review. Simul Healthc. 2021 Aug;16(4):e76-e84.
23. Cook DA, Hatala R, Brydges R, Zendejas B, Hamstra SJ. Technology-Enhanced Simulation for Chronic Disease Management: A Meta-analysis. JAMA Intern Med. 2023 Feb;183(2):123-35.
24. Foronda CL, Fernandez-Burgos M, Nadeau C, Kelley CN, Henry MN. Virtual Simulation in Chronic Disease Education: An Updated Review. Clin Simul Nurs. 2024 Jan;86:101498.
25. Kononowicz AA, Hege I, Edelbring S, Sobocan M. Time-Aware Virtual Patients for Chronic Care Training: A Framework Review. Med Teach. 2023 May;45(5):512-2